A NEW APPROACH TO MUTUAL INFORMATION

FUMIO HIAI [1] AND DÉNES PETZ [2]

Abstract. A new expression as a certain asymptotic limit via "discrete micro-states" of permutations is provided to the mutual information of both continuous and discrete random variables.

Introduction

One of the important quantities in information theory is the mutual information of two random variables X and Y which is expressed in terms of the Boltzmann-Gibbs entropy $H(\cdot)$ as follows:

$$I(X \wedge Y) = -H(X,Y) + H(X) + H(Y)$$

when X, Y are continuous variables. For the expression of $I(X \wedge Y)$ of discrete variables X, Y, the above $H(\cdot)$ is replaced by the Shannon entropy. A more practical and rigorous definition via the relative entropy is

$$I(X \wedge Y) := S(\mu_{(X,Y)}, \mu_X \otimes \mu_Y),$$

where $\mu_{(X,Y)}$ denotes the joint distribution measure of (X,Y) and $\mu_X \otimes \mu_Y$ the product of the respective distribution measures of X, Y.

The aim of this paper is to show that the mutual information $I(X \wedge Y)$ is gained as a certain asymptotic limit of the volume of "discrete micro-states" consisting of permutations approximating joint moments of (X,Y) in some way. In Section 1, more generally we consider an n-tuple of real bounded random variables $(X_1, \dots, X_n)$. Denote by $\Delta(X_1, \dots, X_n; N, m, \delta)$ the set of $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ of $\mathbf{x}_i \in \mathbb{R}^N$ whose joint moments (on the uniform distributed N-point set) of order up to m approximate those of $(X_1, \dots, X_n)$ up to an error δ. Furthermore, denote by $\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, m, \delta)$ the set of $(\sigma_1, \dots, \sigma_n)$ of permutations $\sigma_i \in S_N$ such that $(\sigma_1(\mathbf{x}_1), \dots, \sigma_n(\mathbf{x}_n)) \in \Delta(X_1, \dots, X_n; N, m, \delta)$ for some $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^N_\leq$, where $\mathbb{R}^N_\leq$ is the $\mathbb{R}^N$-vectors arranged in increasing order. Then, the asymptotic volume

$$\frac{1}{N} \log \gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, m, \delta)\bigr)$$

under the uniform probability measure γ_{S_N} on S_N is shown to converge as $\limsup_{N\to\infty}$ (also $\liminf_{N\to\infty}$) and then $\lim_{m\to\infty, \delta\searrow 0}$ to

$$-H(X_1, \dots, X_n) + \sum_{i=1}^n H(X_i)$$

[1] Supported in part by Grant-in-Aid for Scientific Research (B)17340043.
[2] Supported in part by the Hungarian Research Grant OTKA T068258.
AMS subject classification: Primary: 62B10, 94A17.

as long as $H(X_i) > -\infty$ for $1 \le i \le n$. Thus, we obtain a kind of discretization of the mutual information via symmetric group (or permutations).

The approach can be applied to an n-tuple of discrete random variables $(X_1, \dots, X_n)$ as well. But the definition of the Δ_{sym}-set of micro-states for discrete variables is somewhat different from the continuous variable case mentioned above, and we discuss the discrete variable case in Section 2 separately.

The idea comes from the paper [3]. Motivated by theory of mutual free information in [6], a similar approach to Voiculescu's free entropy is provided there. The free entropy is the free probability counterpart of the Boltzmann-Gibbs entropy, and $\mathbb{R}^N$-vectors and the symmetric group S_N here are replaced by Hermitian $N \times N$ matrices and the unitary group $\mathrm{U}(N)$, respectively. In this way, the "discretization approach" here is in some sense a classical analog of the "orbital approach" in [3].

1. The continuous case

For $N \in \mathbb{N}$ let $\mathbb{R}^N_\le$ be the convex cone of the N-dimensional Euclidean space $\mathbb{R}^N$ consisting of $\mathbf{x} = (x_1, \dots, x_N)$ such that $x_1 \le x_2 \le \cdots \le x_N$. The space $\mathbb{R}^N$ is naturally regarded as the real function algebra on the N-point set. Let S_N be the symmetric group of order N (i.e., the permutations on $\{1, 2, \dots, n\}$). Throughout this section let $(X_1, \dots, X_n)$ be an n-tuple of real random variables on a probability space $(\Omega, \mathbb{P})$, and assume that the X_i's are bounded (i.e., $X_i \in L^\infty(\Omega; \mathbb{P})$). The *Boltzmann-Gibbs entropy* of $(X_1, \dots, X_n)$ is defined to be

$$H(X_1, \dots, X_n) := -\int \cdots \int_{\mathbb{R}^n} p(x_1, \dots, x_n) \log p(x_1, \dots, x_n)\, dx_1 \cdots dx_n$$

if the joint density $p(x_1, \dots, x_n)$ of $(X_1, \dots, X_n)$ exists; otherwise $H(X_1, \dots, X_n) = -\infty$. Note that the above integral is well defined in $[-\infty, \infty)$ since the density p is compactly supported.

Definition 1.1. The mean value of $\mathbf{x} = (x_1, \dots, x_N)$ in $\mathbb{R}^N$ is given by

$$\kappa_N(\mathbf{x}) := \frac{1}{N} \sum_{j=1}^N x_j.$$

For each $N, m \in \mathbb{N}$ and $\delta > 0$ we define $\Delta(X_1, \dots, X_n; N, m, \delta)$ to be the set of all n-tuples $(\mathbf{x}_1, \dots, \mathbf{x}_n)$ of $\mathbf{x}_i = (x_{i1}, \dots, x_{iN}) \in \mathbb{R}^N$, $1 \le i \le n$, such that

$$|\kappa_N(\mathbf{x}_{i_1} \cdots \mathbf{x}_{i_k}) - \mathbb{E}(X_{i_1} \cdots X_{i_k})| < \delta$$

for all $1 \le i_1, \dots, i_k \le n$ with $1 \le k \le m$, where $\mathbf{x}_{i_1} \cdots \mathbf{x}_{i_k}$ means the pointwise product, i.e.,

$$\mathbf{x}_{i_1} \cdots \mathbf{x}_{i_k} := (x_{i_1 1} \cdots x_{i_k 1}, x_{i_1 2} \cdots x_{i_k 2}, \dots, x_{i_1 N} \cdots x_{i_k N}) \in \mathbb{R}^N$$

and $\mathbb{E}(\cdot)$ denotes the expectation on $(\Omega, \mathbb{P})$. For each $R > 0$, define $\Delta_R(X_1, \dots, X_n; N, m, \delta)$ to be the set of all $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Delta(X_1, \dots, X_n; N, m, \delta)$ such that $\mathbf{x}_i \in [-R, R]^N$ for all $1 \le i \le n$.

Heuristically, $\Delta(X_1, \dots, X_n; N, m, \delta)$ is the set of "micro-states" consisting of n-tuples of discrete random variables on the N-point set with the uniform probability such that all joint moments of order up to m give the corresponding joint moments of $X_1, \dots, X_n$ up to an error δ.

For $\mathbf{x} \in \mathbb{R}^N$ write $\|\mathbf{x}\|_p := (N^{-1}\sum_{j=1}^N |x_j|^p)^{1/p}$ for $1 \le p < \infty$ and $\|\mathbf{x}\|_\infty := \max_{1\le j\le N} |x_j|$ while $\|X\|_p$ denotes the L^p-norm of a real random variable X on $(\Omega, \mathbb{P})$.

The next lemma is seen from [4, 5.1.1] based on the Sanov large deviation theorem, which says that the Boltzmann-Gibbs entropy is gained as an asymptotic limit of the volume of the approximating micro-states.

Lemma 1.2. *For every $m \in \mathbb{N}$ and $\delta > 0$ and for any choice of $R \ge \max_{1\le i\le n} \|X_i\|_\infty$, the limit*

$$\lim_{N\to\infty} \frac{1}{N} \log \lambda_N^{\otimes n}\big(\Delta_R(X_1, \dots, X_n; N, m, \delta)\big)$$

exists, where λ_N is the Lebesgue measure on $\mathbb{R}^N$. Furthermore, one has

$$H(X_1, \dots, X_n) = \lim_{m\to\infty, \delta\searrow 0} \lim_{N\to\infty} \frac{1}{N} \log \lambda_N^{\otimes n}\big(\Delta_R(X_1, \dots, X_n; N, m, \delta)\big)$$

independently of the choice of $R \ge \max_{1\le i\le n} \|X_i\|_\infty$.

In the following let us introduce some kinds of mutual information in the discretization approach using micro-states of permutations.

Definition 1.3. The action of S_N on $\mathbb{R}^N$ is given by

$$\sigma(\mathbf{x}) := (x_{\sigma^{-1}(1)}, x_{\sigma^{-1}(2)}, \dots, x_{\sigma^{-1}(N)})$$

for $\sigma \in S_N$ and $\mathbf{x} = (x_1, \dots, x_N) \in \mathbb{R}^N$. For each $N, m \in \mathbb{N}$, $\delta > 0$ and $R > 0$ we denote by $\Delta_{\mathrm{sym},R}(X_1, \dots, X_n; N, m, \delta)$ the set of all $(\sigma_1, \dots, \sigma_n) \in S_N^n$ such that

$$(\sigma_1(x_1), \dots, \sigma_n(x_n)) \in \Delta_R(X_1, \dots, X_n; N, m, \delta)$$

for some $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\mathbb{R}^N_\le)^n$. For each $R > 0$ define

$$I_{\mathrm{sym},R}(X_1, \dots, X_n) := -\lim_{m\to\infty, \delta\searrow 0} \limsup_{N\to\infty} \frac{1}{N} \log \gamma_{S_N}^{\otimes n}\big(\Delta_{\mathrm{sym},R}(X_1, \dots, X_n; N, m, \delta)\big),$$

where γ_{S_N} is the uniform probability measure on S_N. Define also $\overline{I}_{\mathrm{sym},R}(X_1, \dots, X_n)$ by replacing lim sup by lim inf. Obviously,

$$0 \le I_{\mathrm{sym},R}(X_1, \dots, X_n) \le \overline{I}_{\mathrm{sym},R}(X_1, \dots, X_n).$$

Moreover, $\Delta_{\mathrm{sym},\infty}(X_1, \dots, X_n; N, m, \delta)$ is defined by replacing $\Delta_R(X_1, \dots, X_n; N, m, \delta)$ in the above by $\Delta(X_1, \dots, X_n; N, m, \delta)$ without cut-off by the parameter R. Then $I_{\mathrm{sym},\infty}(X_1, \dots, X_n)$ and $\overline{I}_{\mathrm{sym},\infty}(X_1, \dots, X_n)$ are also defined as above.

Definition 1.4. For each $1 \le i \le n$ we choose and fix a sequence $\xi_i = \{\xi_i(N)\}$ of $\xi_i(N) \in \mathbb{R}^N_\le$, $N \in \mathbb{N}$, such that $\kappa_N(\xi_i(N)^k) \to \mathbb{E}(X_i^k)$ as $N \to \infty$ for all $k \in \mathbb{N}$, i.e., $\xi_i(N) \to X_i$ in moments. For each $N, m \in \mathbb{N}$ and $\delta > 0$ we define $\Delta_{\mathrm{sym}}(X_1, \dots, X_n : \xi_1(N), \dots, \xi_n(N); N, m, \delta)$ to be the set of all $(\sigma_1, \dots, \sigma_n) \in S_N^n$ such that

$$(\sigma_1(\xi_1(N)), \dots, \sigma_n(\xi_n(N))) \in \Delta(X_1, \dots, X_n; N, m, \delta).$$

Define

$$
\begin{aligned}
&I_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n)\\
&\quad:= -\lim_{m\to\infty,\delta\searrow 0}\limsup_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta)\bigr)
\end{aligned}
$$

and $\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots\xi_n)$ by replacing lim sup by lim inf.

The next proposition asserts that the quantities in Definitions 1.3 and 1.4 are all equivalent.

Lemma 1.5. *For any choice of* $R\ge\max_{1\le i\le n}\|X_i\|_\infty$ *and for any choices of approximating sequences* $\xi_1,\dots,\xi_n$ *one has*

$$
I_{\mathrm{sym},\infty}(X_1,\dots,X_n)=I_{\mathrm{sym},R}(X_1,\dots,X_n)=I_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n), \tag{1.1}
$$

$$
\overline{I}_{\mathrm{sym},\infty}(X_1,\dots,X_n)=\overline{I}_{\mathrm{sym},R}(X_1,\dots,X_n)=\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n). \tag{1.2}
$$

Proof. It is obvious that $\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta)$ is included in $\Delta_{\mathrm{sym},\infty}(X_1,\dots,X_n;N,m,\delta)$ for any approximating sequences ξ_i. Moreover, for each $1\le i\le n$ an approximating sequence ξ_i can be chosen so that $\|\xi_i(N)\|_\infty\le\|X_i\|_\infty$ for all N; then $\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta)\subset\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta)$ for any $R\ge R_0:=\max_{1\le i\le n}\|X_i\|_\infty$. Hence it suffices to prove that for any approximating sequences ξ_i and for every $m\in\mathbb{N}$ and $\delta>0$, there are an $m'\in\mathbb{N}$, a $\delta'>0$ and an $N_0\in\mathbb{N}$ so that

$$
\Delta_{\mathrm{sym},\infty}(X_1,\dots,X_n;N,m',\delta')\subset\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta)
$$

for all $N\ge N_0$. Choose a $\rho\in(0,1)$ with $m(R_0+1)^{m-1}\rho<\delta/2$. By [5, Lemma 4.3] (also [4, 4.3.4]) there exist an $m'\in\mathbb{N}$ with $m'\ge 2m$, a $\delta'>0$ with $\delta'\le\min\{1,\delta/2\}$ and an $N_0\in\mathbb{N}$ such that for every $1\le i\le n$ and every $\mathbf{x}\in\mathbb{R}^N_\le$ with $N\ge N_0$, if $|\kappa_N(\mathbf{x}^k)-\mathbb{E}(X_i^k)|<\delta'$ for all $1\le k\le m'$, then $\|\mathbf{x}-\xi_i(N)\|_m<\rho$. Suppose $N\ge N_0$ and $(\sigma_1,\dots,\sigma_n)\in\Delta_{\mathrm{sym},\infty}(X_1,\dots,X_n;N,m',\delta')$; then $(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))\in\Delta(X_1,\dots,X_n;N,m',\delta')$ for some $(\mathbf{x}_1,\dots,\mathbf{x}_n)\in(\mathbb{R}^N_\le)^n$. Since $|\kappa_N(\mathbf{x}_i^k)-\mathbb{E}(X_i^k)|<\delta'$ for all $1\le k\le m'$, we get $\|\mathbf{x}_i-\xi_i(N)\|_m\le\rho$ and

$$
\begin{aligned}
\|\mathbf{x}_i\|_m\le\|\mathbf{x}_i\|_{2m}&=\kappa_N(\mathbf{x}_i^{2m})^{1/2m}\\
&<(\mathbb{E}(X_i^{2m})+1)^{1/2m}\\
&\le(R_0^{2m}+1)^{1/2m}\le R_0+1.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&|\kappa_N(\sigma_{i_1}(\xi_{i_1}(N))\cdots\sigma_{i_k}(\xi_{i_k}(N)))-\mathbb{E}(X_{i_1}\cdots X_{i_k})|\\
&\quad\le|\kappa_N(\sigma_{i_1}(\xi_{i_1}(N))\cdots\sigma_{i_k}(\xi_{i_k}(N)))-\kappa_N(\sigma_{i_1}(\mathbf{x}_{i_1})\cdots\sigma_{i_k}(\mathbf{x}_{i_k}))|\\
&\qquad+|\kappa_N(\sigma_{i_1}(\mathbf{x}_{i_1})\cdots\sigma_{i_k}(\mathbf{x}_{i_k}))-\mathbb{E}(X_{i_1}\cdots X_{i_k})|\\
&\quad\le m(R_0+1)^{m-1}\rho+\delta'<\delta
\end{aligned}
$$

for all $1\le i_1,\dots,i_k\le n$ with $1\le k\le m$. The above latter inequality follows from the Hölder inequality. Hence $(\sigma_1,\dots,\sigma_n)\in\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta)$, and the result follows. □

Consequently, we denote all the quantities in (1.1) by the same $I_{\mathrm{sym}}(X_1,\dots,X_n)$ and those in (1.2) by $\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)$. We call $I_{\mathrm{sym}}(X_1,\dots,X_n)$ and $\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)$ the *mutual information* and *upper mutual information* of $(X_1,\dots,X_n)$, respectively. The terminology "mutual information" will be justified after the next theorem.

In the continuous variable case, our main result is the following exact relation of I_{sym} and $\overline{I}_{\mathrm{sym}}$ with the Boltzmann-Gibbs entropy $H(\cdot)$, which says that $I_{\mathrm{sym}}(X_1,\dots,X_n)$ is formally the sum of the separate entropies $H(X_i)$'s minus the compound $H(X_1,\dots,X_n)$. Thus, a naive meaning of $I_{\mathrm{sym}}(X_1,\dots,X_n)$ is the entropy (or information) overlapping among the X_i's.

Theorem 1.6.

$$
\begin{aligned}
H(X_1,\dots,X_n) &= -I_{\mathrm{sym}}(X_1,\dots,X_n) + \sum_{i=1}^n H(X_i) \\
&= -\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n) + \sum_{i=1}^n H(X_i).
\end{aligned}
$$

Proof. If the coordinates s_i of $\mathbf{s}\in\mathbb{R}^N$ are all distinct, then $\mathbf{s}$ is uniquely written as $\mathbf{s}=\sigma(\mathbf{x})$ with $\mathbf{x}\in\mathbb{R}^N_\le$ and $\sigma\in S_N$. Note that the set of $\mathbf{s}\in\mathbb{R}^N$ with $s_i=s_j$ for some $i\ne j$ is a closed subset of λ_N-measure zero. Under the correspondence

$$
\mathbf{s}\in\mathbb{R}^N \longleftrightarrow (\mathbf{x},\sigma)\in\mathbb{R}^N_\le\times S_N, \quad \mathbf{s}=\sigma(\mathbf{x})
$$

(well defined on a co-negligible subset of $\mathbb{R}^N$), the measure λ_N is transformed into the product of $\lambda_N|_{\mathbb{R}^N_\le}$ and the counting measure on S_N.

In the following proof we adopt, due to Lemma 1.5, the description of I_{sym} and $\overline{I}_{\mathrm{sym}}$ as $I_{\mathrm{sym},R}(X_1,\dots,X_n)$ and $\overline{I}_{\mathrm{sym},R}(X_1,\dots,X_n)$ with $R:=\max_{1\le i\le n}\|X_i\|_\infty$. For each $N,m\in\mathbb{N}$ and $\delta>0$, suppose $(\mathbf{s}_1,\dots,\mathbf{s}_n)\in\Delta_R(X_1,\dots,X_n;N,m,\delta)$ and write $\mathbf{s}_i=\sigma_i(\mathbf{x}_i)$ with $\mathbf{x}_i\in\mathbb{R}^N_\le$ and $\sigma_i\in S_N$. Then it is obvious that

$$
\begin{aligned}
&(\mathbf{x}_1,\dots,\mathbf{x}_n;\sigma_1,\dots,\sigma_n) \\
&\quad\in \left(\prod_{i=1}^n \bigl(\Delta_R(X_i;N,m,\delta)\cap\mathbb{R}^N_\le\bigr)\right)\times\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta).
\end{aligned}
$$

By Lemma 1.2 and the fact stated at the beginning of the proof, we obtain

$$
\begin{aligned}
H(X_1,\dots,X_n) &\le \lim_{N\to\infty}\frac{1}{N}\log\lambda_N^{\otimes n}\bigl(\Delta_R(X_1,\dots,X_n;N,m,\delta)\bigr) \\
&\le \liminf_{N\to\infty}\frac{1}{N}\left(\sum_{i=1}^n \log\lambda_N\bigl(\Delta_R(X_i;N,m,\delta)\cap\mathbb{R}^N_\le\bigr)\right. \\
&\qquad\qquad \left. + \log\#\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta)\right) \\
&= \liminf_{N\to\infty}\frac{1}{N}\left(\sum_{i=1}^n \log\lambda_N\bigl(\Delta_R(X_i;N,m,\delta)\bigr) - n\log N!\right.
\end{aligned}
$$

$$
\begin{aligned}
&\qquad\qquad + \log \#\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta)\Bigg) \\
&= \sum_{i=1}^n \lim_{N\to\infty} \frac{1}{N}\log\lambda_N\bigl(\Delta_R(X_i;N,m,\delta)\bigr) \\
&\qquad + \liminf_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta)\bigr).
\end{aligned}
$$

This implies that

$$
H(X_1,\dots,X_n) \le \sum_{i=1}^n H(X_i) - \overline{I}_{\mathrm{sym}}(X_1,\dots,X_n). \tag{1.3}
$$

Conversely, for each $m\in\mathbb{N}$ and $\delta>0$, by [5, Lemma 4.3] (also [4, 4.3.4]) there are an $m'\in\mathbb{N}$ with $m'\ge m$, a $\delta'>0$ with $\delta'\le\delta/2$ and an $N_0\in\mathbb{N}$ such that for every $N\in\mathbb{N}$ and for every $\mathbf{x},\mathbf{y}\in\mathbb{R}_\le^N$, if $\|\mathbf{x}\|_\infty\le R$ and $|\kappa_N(\mathbf{x}^k)-\kappa_N(\mathbf{y}^k)|<2\delta'$ for all $1\le k\le m'$, then $\|\mathbf{x}-\mathbf{y}\|_1<\delta/2m(R+1)^{m-1}$. Suppose $N\ge N_0$ and

$$
\begin{aligned}
&(\mathbf{x}_1,\dots,\mathbf{x}_n;\sigma_1,\dots,\sigma_n) \\
&\quad\in\left(\prod_{i=1}^n\bigl(\Delta_R(X_i;N,m',\delta')\cap\mathbb{R}_\le^N\bigr)\right)\times\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m',\delta')
\end{aligned}
$$

so that $(\sigma_1(\mathbf{y}_1),\dots,\sigma_n(\mathbf{y}_n))\in\Delta_R(X_1,\dots,X_n;N,m',\delta')$ for some $(\mathbf{y}_1,\dots,\mathbf{y}_n)\in(\mathbb{R}_\le^N)^n$. Since

$$
|\kappa_N(\mathbf{x}_i^k)-\kappa_N(\mathbf{y}_i^k)|\le|\kappa_N(\mathbf{x}_i^k)-\mathbb{E}(X_i^k)|+|\kappa_N(\mathbf{y}_i^k)-\mathbb{E}(X_i^k)|<2\delta'
$$

for all $1\le k\le m'$, we get $\|\mathbf{x}_i-\mathbf{y}_i\|_1<\delta/2m(R+1)^{m-1}$ for $1\le i\le n$. Therefore,

$$
\begin{aligned}
&|\kappa_N(\sigma_{i_1}(\mathbf{x}_{i_1})\cdots\sigma_{i_k}(\mathbf{x}_{i_k}))-\mathbb{E}(X_{i_1}\cdots X_{i_k})| \\
&\quad\le|\kappa_N(\sigma_{i_1}(\mathbf{x}_{i_1})\cdots\sigma_{i_k}(\mathbf{x}_{i_k}))-\kappa_N(\sigma_{i_1}(\mathbf{y}_{i_1})\cdots\sigma_{i_k}(\mathbf{y}_{i_k}))| \\
&\qquad+|\kappa_N(\sigma_{i_1}(\mathbf{y}_{i_1})\cdots\sigma_{i_k}(\mathbf{y}_{i_k}))-\mathbb{E}(X_{i_1}\cdots X_{i_k})| \\
&\quad\le m(R+1)^{m-1}\max_{1\le i\le n}\|\mathbf{x}_i-\mathbf{y}_i\|_1+\delta' \\
&\quad<\frac{\delta}{2}+\delta'\le\delta
\end{aligned}
$$

for all $1\le i_1,\dots,i_k\le n$ with $1\le k\le m$. This implies that $(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))\in\Delta_R(X_1,\dots,X_n;N,m,\delta)$. By Lemma 1.2 we obtain

$$
\begin{aligned}
&\sum_{i=1}^n H(X_i)-I_{\mathrm{sym}}(X_1,\dots,X_n) \\
&\quad\le\sum_{i=1}^n\lim_{N\to\infty}\frac{1}{N}\log\lambda_N\bigl(\Delta_R(X_i;N,m',\delta')\bigr) \\
&\qquad+\limsup_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m',\delta')\bigr) \\
&\quad=\limsup_{N\to\infty}\frac{1}{N}\left(\sum_{i=1}^n\log\lambda_N\bigl(\Delta_R(X_i;N,m',\delta')\cap\mathbb{R}_\le^N\bigr)\right.
\end{aligned}
$$

$$+ \log \#\Delta_{\mathrm{sym},R}(X_1, \dots, X_n; N, m', \delta')\Bigg)$$

$$\le \limsup_{N\to\infty} \frac{1}{N} \log \lambda_N^{\otimes n}\big(\Delta_R(X_1, \dots, X_n; N, m, \delta)\big).$$

This implies by Lemma 1.2 once again that

$$\sum_{i=1}^{n} H(X_i) - I_{\mathrm{sym}}(X_1, \dots, X_n) \le H(X_1, \dots, X_n). \tag{1.4}$$

The result follows from (1.3) and (1.4). □

Let $\mu_{(X_1,\dots,X_n)}$ be the joint distribution measure on $\mathbb{R}^n$ of $(X_1, \dots, X_n)$ while μ_{X_i} is that of X_i for $1 \le i \le n$. Let $S(\mu_{(X_1,\dots,X_n)}, \mu_{X_1} \otimes \cdots \otimes \mu_{X_n})$ denote the *relative entropy* (or *the Kullback-Leibler divergence*) of $\mu_{(X_1,\dots,X_n)}$ with respect to the product measure $\mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$, i.e.,

$$S(\mu_{(X_1,\dots,X_n)}, \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}) := \int \log \frac{d\mu_{(X_1,\dots,X_n)}}{d(\mu_{X_1} \otimes \cdots \otimes \mu_{X_n})}\, d\mu_{(X_1,\dots,X_n)}$$

if $\mu_{(X_1,\dots,X_n)}$ is absolutely continuous with respect to $\mu_{X_1} \otimes \cdots \otimes \mu_{X_n}$; otherwise $S(\mu_{(X_1,\dots,X_n)}, \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}) := +\infty$. When $H(X_i) > -\infty$ for all $1 \le i \le n$, one can easily verify that

$$S(\mu_{(X_1,\dots,X_n)}, \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}) = -H(X_1, \dots, X_n) + \sum_{i=1}^{n} H(X_i).$$

Thus, the above theorem yields the following:

Corollary 1.7. *If $H(X_i) > -\infty$ for all $1 \le i \le n$, then*

$$\begin{aligned} I_{\mathrm{sym}}(X_1, \dots, X_n) &= \overline{I}_{\mathrm{sym}}(X_1, \dots, X_n) \\ &= S(\mu_{(X_1,\dots,X_n)}, \mu_{X_1} \otimes \cdots \otimes \mu_{X_n}). \end{aligned}$$

Corollary 1.8. *Under the same assumption as the above corollary, $I_{\mathrm{sym}}(X_1, \dots, X_n) = 0$ if and only if $X_1, \dots, X_n$ are independent.*

In particular, the original *mutual information* $I(X_1 \wedge X_2)$ of two real random variables X_1, X_2 is normally defined as

$$I(X_1 \wedge X_2) := S(\mu_{(X_1,X_2)}, \mu_{X_1} \otimes \mu_{X_2}).$$

Hence we have

$$I(X_1 \wedge X_2) = I_{\mathrm{sym}}(X_1, X_2) = \overline{I}_{\mathrm{sym}}(X_1, X_2)$$

as long as $H(X_1) > -\infty$ and $H(X_2) > -\infty$ (and X_1, X_2 are bounded). For this reason, we gave the term "mutual information" to I_{sym}.

Finally, some open problems are in order:

(1) Without the assumption $H(X_i) > -\infty$ for $1 \le i \le n$, does $I_{\mathrm{sym}}(X_1, \dots, X_n) = \overline{I}_{\mathrm{sym}}(X_1, \dots, X_n)$ hold true?

(2) More strongly, does the limit such as

$$\lim_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}(\Delta_{\mathrm{sym},R}(X_1,\dots,X_n;N,m,\delta))$$

or

$$\lim_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}(\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,m,\delta))$$

exist as in Lemma 1.2?

(3) Without the assumption $H(X_i)>-\infty$ for $1\le i\le n$, does $I_{\mathrm{sym}}(X_1,\dots,X_n)=S(\mu_{(X_1,\dots,X_n)},\mu_{X_1}\otimes\cdots\otimes\mu_{X_n})$ hold true? Also, is $I_{\mathrm{sym}}(X_1,\dots,X_n)=0$ equivalent to the independence of $X_1,\dots,X_n$?

(4) Although the boundedness assumption for $X_1,\dots,X_n$ is rather essential in the above discussions, it is desirable to extend the results in this section to $X_1,\dots,X_n$ not necessarily bounded but having all moments.

2. The discrete case

Let $\mathcal{Y}$ be a finite set with a probability measure p. The *Shannon entropy* of p is

$$S(p):=-\sum_{y\in\mathcal{Y}}p(y)\log p(y).$$

For each sequence $\mathbf{y}=(y_1,\dots,y_N)\in\mathcal{Y}^N$, the *type* of $\mathbf{y}$ is a probability measure on $\mathcal{Y}$ given by

$$\nu_{\mathbf{y}}(t):=\frac{N_{\mathbf{y}}(t)}{N}\quad\text{where}\quad N_{\mathbf{y}}(t):=\#\{j:y_j=t\},\quad t\in\mathcal{Y}.$$

The number of possible types is smaller than $(N+1)^{\#\mathcal{Y}}$. If ν is a type and $\mathcal{T}_N(\nu)$ denotes the set of all sequences of type ν from $\mathcal{Y}^N$, then the cardinality of $\mathcal{T}_N(\nu)$ is estimated as follows:

$$\frac{1}{(N+1)^{\#\mathcal{Y}}}\,e^{NS(\nu)}\le\#\mathcal{T}_N(\nu)\le e^{NS(\nu)}\tag{2.1}$$

(see [1, 12.1.3] and [2, Lemma 2.2]).

Let p be a probability meausre on $\mathcal{Y}$. For each $N\in\mathbb{N}$ and $\delta>0$ we define $\Delta(p;N,\delta)$ to be the set of all sequences $\mathbf{y}\in\mathcal{Y}^N$ such that $|\nu_{\mathbf{y}}(t)-p(t)|<\delta$ for all $t\in\mathcal{Y}$. In other words, $\Delta(p;N,\delta)$ is the set of all δ-typical sequeces (with respect to the measure p). Then the next lemma is well known.

Lemma 2.1.

$$S(p)=\lim_{\delta\searrow0}\lim_{N\to\infty}\frac{1}{N}\log\#\Delta(p;N,\delta).$$

In fact, this easily follows from (2.1). Let $P_{N,\delta}$ be the maximizer of the Shannon entropy on the set of all types $\nu_{\mathbf{y}}$, $\mathbf{y}\in\mathcal{Y}^N$, such that $|\nu_{\mathbf{y}}(t)-p(t)|<\delta$ for all $t\in\mathcal{Y}$. We can use the Shannon entropy of the type class corresponding to $P_{N,\delta}$ to estimate the cardinality of $\Delta(p;N,\delta)$:

$$(N+1)^{-\#\mathcal{Y}}e^{NS(P_{N,\delta})}\le\#\Delta(p;N,\delta)\le e^{NS(P_{N,\delta})}(N+1)^{\#\mathcal{Y}}.$$

It follows that

$$\lim_{N\to\infty}\frac{1}{N}\log\#\Delta(p;N,\delta)=\sup\{S(q):q\text{ is a probability meausre on }\mathcal{Y}\\ \text{such that }|q(t)-p(t)|<\delta,\,t\in\mathcal{Y}\},$$

and the lemma follows.

We consider the case where p is the joint distribution of an n-tuple $(X_1,\dots,X_n)$ of discrete random variables on $(\Omega,\mathbb{P})$. Throughout this section we assume that the random variables $X_1,\dots,X_n$ have their values in a finite set $\mathcal{X}=\{t_1,\dots,t_d\}$.

Definition 2.2. Let $p_{(X_1,\dots,X_n)}$ denote the joint distribution of $(X_1,\dots,X_n)$, which is a measure on $\mathcal{X}^n$ while the distribution p_{X_i} of X_i is a measure on $\mathcal{X}$, $1\le i\le n$. We write $\Delta(X_i;N,\delta)$ for $\Delta(p_{X_i};N,\delta)$ and $\Delta(X_1,\dots,X_n;N,\delta)$ for $\Delta(p_{(X_1,\dots,X_n)};N,\delta)$.

Next, we introduce the counterparts of Definitions 1.3 and 1.4 in the discrete variable case.

Definition 2.3. The action of S_N on $\mathcal{X}^N$ is similar to that on $\mathbb{R}^N$ given in Defintion 1.3. For $N\in\mathbb{N}$ let $\mathcal{X}^N_\le$ denote the set of all sequences of length N of the form

$$\mathbf{x}=(t_1,\dots,t_1,t_2,\dots,t_2,\dots,t_d,\dots,t_d).$$

Oviously, such a sequence $\mathbf{x}$ is uniquely determined by $(N_{\mathbf{x}}(t_1),\dots,N_{\mathbf{x}}(t_d))$ or the type of $\mathbf{x}$. That is, $\mathcal{X}^N_\le$ is regarded as the set of all types from $\mathcal{X}^N$. For each $N\in\mathbb{N}$ and $\delta>0$ we denote by $\Delta_{\mathrm{sym}}(X_1,\dots,X_n;N,\delta)$ the set of all $(\sigma_1,\dots,\sigma_n)\in S_N^n$ such that

$$(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))\in\Delta(X_1,\dots,X_n;N,\delta)$$

for some $(\mathbf{x}_1,\dots,\mathbf{x}_n)\in(\mathcal{X}^N_\le)^n$. Define

$$I_{\mathrm{sym}}(X_1,\dots,X_n):=-\lim_{\delta\searrow0}\limsup_{N\to\infty}\frac{1}{N}\log\gamma_{S_N}^{\otimes n}(\Delta_{\mathrm{sym}}(X_1,\dots,X_n;N,\delta)),$$

and $\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)$ by replacing lim sup by lim inf. Moreover, for each $1\le i\le n$, choose a sequence $\xi_i=\{\xi_i(N)\}$ of $\xi_i(N)=(\xi_i(N)_1,\dots,\xi_i(N)_N)\in\mathcal{X}^N_\le$ such that $\nu_{\xi_i(N)}\to p_{X_i}$ as $N\to\infty$. We then define $\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,\delta)$, $I_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n)$ and $\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n)$ as in Definition 1.4.

Lemma 2.4. *For any choices of approximating sequences $\xi_1,\dots,\xi_n$ one has*

$$I_{\mathrm{sym}}(X_1,\dots,X_n)=I_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n),$$
$$\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)=\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1,\dots,\xi_n).$$

Proof. It suffices to show that for each $\delta>0$ there are a $\delta'>0$ and an $N_0\in\mathbb{N}$ such that

$$\Delta_{\mathrm{sym}}(X_1,\dots,X_n;N,\delta')\subset\Delta_{\mathrm{sym}}(X_1,\dots,X_n:\xi_1(N),\dots,\xi_n(N);N,\delta)\tag{2.2}$$

for all $N\ge N_0$. Choose $\delta'>0$ so that $3nd^{n+1}\delta'\le\delta$, where $d=\#\mathcal{X}$. Suppose $(\sigma_1,\dots,\sigma_n)$ is in the left-hand side of (2.2) so that $(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))\in\Delta(X_1,\dots,X_n;N,\delta')$ for some $(\mathbf{x}_1,\dots,\mathbf{x}_n)$, $\mathbf{x}_i=(x_{i1},\dots,x_{iN})\in\mathcal{X}^N_\le$. Since

$$|\nu_{(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))}(z_1,\dots,z_n)-p_{(X_1,\dots,X_n)}(z_1,\dots,z_n)|<\delta',\quad(z_1,\dots,z_n)\in\mathcal{X}^n,\tag{2.3}$$

$$\nu_{\mathbf{x}_i}(t) = \sum_{z_1,\dots,z_{i-1},z_{i+1},\dots,z_n\in\mathcal{X}} \nu_{(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))}(z_1,\dots,z_{i-1},t,z_{i+1},\dots,z_n), \quad t\in\mathcal{X},$$

$$p_{X_i}(t) = \sum_{z_1,\dots,z_{i-1},z_{i+1},\dots,z_n\in\mathcal{X}} p_{(X_1,\dots,X_n)}(z_1,\dots,z_{i-1},t,z_{i+1},\dots,z_n), \quad t\in\mathcal{X},$$

it follows that

$$|\nu_{\mathbf{x}_i}(t) - p_{X_i}(t)| < d^{n-1}\delta' \tag{2.4}$$

for any $1\le i\le n$ and $t\in\mathcal{X}$. Now, choose an $N_0\in\mathbb{N}$ so that $|\nu_{\xi_i(N)}(t)-p_{X_i}(t)|<\delta'$ and hence

$$|\nu_{\xi_i(N)}(t) - \nu_{\mathbf{x}_i}(t)| < 2d^{n-1}\delta' \tag{2.5}$$

for any $1\le i\le n$ and $t\in\mathcal{X}$ and for all $N\ge N_0$. Since

$$\begin{aligned} &|(N_{\xi_i(N)}(t_1)+\cdots+N_{\xi_i(N)}(t_l)) - (N_{\mathbf{x}_i}(t_1)+\cdots+N_{\mathbf{x}_i}(t_l))| \\ &\quad\le |N_{\xi_i(N)}(t_1)-N_{\mathbf{x}_i}(t_1)|+\cdots+|N_{\xi_i(N)}(t_l)-N_{\mathbf{x}_i}(t_l)| \\ &\quad< 2Nd^n\delta' \end{aligned}$$

for every $1\le l\le d$ thanks to (2.5), it is easily seen that

$$\#\big\{j\in\{1,\dots,N\} : \xi_i(N)_j \ne x_{ij}\big\} < 2Nd^{n+1}\delta'$$

for any $1\le i\le n$. Hence we get

$$\begin{aligned} &|\nu_{(\sigma_1(\xi_1(N)),\dots,\sigma_n(\xi_n(N)))}(z_1,\dots,z_n) - \nu_{(\sigma_1(\mathbf{x}_1),\dots,\sigma_n(\mathbf{x}_n))}(z_1,\dots,z_n)| \\ &\quad= \frac{1}{N}\big|\#\{j : \xi_1(N)_{\sigma_1^{-1}(j)}=z_1,\dots,\xi_n(N)_{\sigma_n^{-1}(j)}=z_n\} \\ &\qquad\qquad - \#\{j : x_{1\sigma_1^{-1}(j)}=z_1,\dots,x_{n\sigma_n^{-1}(j)}=z_n\}\big| \\ &\quad\le \frac{1}{N}\sum_{i=1}^n \#\{j:\xi_i(N)_j\ne x_{ij}\} < 2nd^{n+1}\delta' \end{aligned}$$

so that thanks to (2.3)

$$|\nu_{(\sigma_1(\xi_1(N)),\dots,\sigma_n(\xi_n(N)))}(z_1,\dots,z_n) - p_{(X_1,\dots,X_n)}(z_1,\dots,z_n)| < 3nd^{n+1}\delta' \le \delta$$

for every $(z_1,\dots,z_n)\in\mathcal{X}^n$. Therefore, $(\sigma_1,\dots,\sigma_n)$ is in the right-hand side of (2.2), as required. □

The next theorem is the discrete variable version of Theorem 1.6.

Theorem 2.5.

$$I_{\mathrm{sym}}(X_1,\dots,X_n) = \overline{I}_{\mathrm{sym}}(X_1,\dots,X_n) = -S(X_1,\dots,X_n) + \sum_{i=1}^n S(X_i).$$

Proof. For each sequence $(N_1,\dots,N_d)$ of integers $N_l\ge 0$ with $\sum_{l=1}^d N_l = N$, let $S(N_1,\dots,N_d)$ denote the subgroup of S_N consisting of products of permutations of $\{1,\dots,N_1\}$, $\{N_1+1,\dots,N_1+N_2\}$, …, $\{N_1+\cdots+N_{d-1}+1,\dots,N\}$, and let

$$S_N/S(N_1,\dots,N_d)$$

be the set of left cosets of $S(N_1,\dots,N_d)$. For each $\mathbf{x}\in\mathcal{X}_\le^N$ and $\sigma\in S_N$ we write $[\sigma]_{\mathbf{x}}$ for the left coset of $S(N_{\mathbf{x}}(t_1),\dots,N_{\mathbf{x}}(t_d))$ containing σ. Then it is clear that every $\mathbf{s} \in \mathcal{X}^N$ is represented as $\mathbf{s} = \sigma(\mathbf{x})$ with a unique pair $(\mathbf{x}, [\sigma]_\mathbf{x})$ of $\mathbf{x} \in \mathcal{X}_\leq^N$ and $[\sigma]_\mathbf{x} \in S_N/S(N_\mathbf{x}(t_1), \dots, N_\mathbf{x}(t_d))$.

For any $\varepsilon > 0$ one can choose a $\delta > 0$ such that for every $1 \leq i \leq n$ and every probability measure p on $\mathcal{X}$, if $|p(t) - p_{X_i}(t)| < \delta$ for all $t \in \mathcal{X}$, then $|S(p) - S(p_{X_i})| < \varepsilon$. This implies that for each $N \in \mathbb{N}$ and $1 \leq i \leq n$, one has $|S(\nu_\mathbf{x}) - S(p_{X_i})| < \varepsilon$ whenever $\mathbf{x} \in \Delta(X_i; N, \delta)$. Notice that $\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, \delta/d^{n-1})$ is the union of $[\sigma_1]_{\mathbf{x}_1} \times \cdots \times [\sigma_n]_{\mathbf{x}_n}$ for all $(\mathbf{x}_1, \dots, \mathbf{x}_n; [\sigma_1]_{\mathbf{x}_1}, \dots, [\sigma_n]_{\mathbf{x}_n})$ of $\mathbf{x}_i \in \mathcal{X}_\leq^N$ and $[\sigma_i]_{\mathbf{x}_i} \in S_N/S(N_{\mathbf{x}_i}(t_1), \dots, N_{\mathbf{x}_i}(t_d))$ such that $(\sigma_1(\mathbf{x}_1), \dots, \sigma_n(\mathbf{x}_n)) \in \Delta(X_1, \dots, X_n; N, \delta/d^{n-1})$. Now, suppose $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\mathcal{X}_\leq^N)^n$, $(\sigma_1, \dots, \sigma_n) \in S_N^n$ and $(\sigma_1(\mathbf{x}_1), \dots, \sigma_n(\mathbf{x}_n)) \in \Delta(X_1, \dots, X_n; N, \delta/d^{n-1})$. Then, for each $1 \leq i \leq n$ we get $\mathbf{x}_i \in \Delta(X_i; N, \delta)$, i.e., $|\nu_{\mathbf{x}_i}(t) - p_{X_i}(t)| < \delta$ for all $t \in \mathcal{X}$ as (2.4). Hence we have

$$\#\big([\sigma_1]_{\mathbf{x}_1} \times \cdots \times [\sigma_n]_{\mathbf{x}_n}\big) \leq \prod_{i=1}^n \left(\max_{\mathbf{x} \in \Delta(X_i; N, \delta)} \prod_{t \in \mathcal{X}} N_\mathbf{x}(t)! \right) \tag{2.6}$$

so that

$$\begin{aligned} &\#\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, \delta/d^{n-1}) \\ &\quad \leq \#\Delta(X_1, \dots, X_n; N, \delta/d^{n-1}) \cdot \prod_{i=1}^n \left(\max_{\mathbf{x} \in \Delta(X_i; N, \delta)} \prod_{t \in \mathcal{X}} N_\mathbf{x}(t)! \right). \end{aligned}$$

Therefore,

$$\begin{aligned} &\frac{1}{N} \log \gamma_{S_N}^{\otimes n}\big(\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, \delta/d^{n-1})\big) \\ &\quad \leq \frac{1}{N} \log \#\Delta(X_1, \dots, X_n; N, \delta/d^{n-1}) \\ &\qquad + \sum_{i=1}^n \max_{\mathbf{x} \in \Delta(X_i; N, \delta)} \left(\frac{1}{N} \sum_{t \in \mathcal{X}} \log N_\mathbf{x}(t)! \right) - \frac{n}{N} \log N!. \end{aligned} \tag{2.7}$$

For each $1 \leq i \leq n$ and for any $\mathbf{x} \in \Delta(X_i; N, , \delta)$, the Stirling formula yields

$$\begin{aligned} &\frac{1}{N} \sum_{t \in \mathcal{X}} \log N_\mathbf{x}(t)! - \frac{1}{N} \log N! \\ &\quad = \sum_{t \in \mathcal{X}} \left(\frac{N_\mathbf{x}(t)}{N} \log N_\mathbf{x}(t) - \frac{N_\mathbf{x}(t)}{N} \right) - \log N + 1 + o(1) \\ &\quad = -S(\nu_\mathbf{x}) + o(1) \leq -S(p_{X_i}) + \varepsilon + o(1) \quad \text{as } N \to \infty \end{aligned} \tag{2.8}$$

thanks to the above choice of $\delta > 0$. Here, note that the $o(1)$ in the above estimate is uniform for $\mathbf{x} \in \Delta(X_i; N, \delta)$. Hence, by (2.7), (2.8) and by Lemma 2.1 applied to $p_{(X_1, \dots, X_n)}$ on $\mathcal{X}^n$, we obtain

$$-I_{\mathrm{sym}}(X_1, \dots, X_n) \leq S(p_{(X_1, \dots, X_n)}) - \sum_{i=1}^n S(p_{X_i}) + n\varepsilon$$

and hence

$$I_{\mathrm{sym}}(X_1, \dots, X_n) \geq -S(X_1, \dots, X_n) + \sum_{i=1}^n S(X_i). \tag{2.9}$$

Next, we prove the converse direction. For any $\varepsilon > 0$ choose a $\delta > 0$ as above. For $N \in \mathbb{N}$ let $\Xi(N, \delta/d^{n-1})$ be the set of all $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in (\mathcal{X}_{\le}^N)^n$ such that

$$(\sigma_1(\mathbf{x}_1), \dots, \sigma_n(\mathbf{x}_n)) \in \Delta(X_1, \dots, X_n; N, \delta/d^{n-1})$$

for some $(\sigma_1, \dots, \sigma_n) \in S_N^n$. Furthermore, for each $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Xi(N, \delta/d^{n-1})$, let $\Sigma(\mathbf{x}_1, \dots, \mathbf{x}_n; N, \delta/d^{n-1})$ be the set of all

$$([\sigma_1]_{\mathbf{x}_1}, \dots, [\sigma_n]_{\mathbf{x}_n}) \in \prod_{i=1}^n S_N/S(N_{\mathbf{x}_i}(t_1), \dots, N_{\mathbf{x}_i}(t_d))$$

such that $(\sigma_1(\mathbf{x}_1), \dots, \sigma_n(\mathbf{x}_n)) \in \Delta(X_1, \dots, X_n; N, \delta/d^{n-1})$. Then it is obvious that

$$\#\Delta(X_1, \dots, X_n; N, \delta/d^{n-1}) \le \sum_{(\mathbf{x}_1,\dots,\mathbf{x}_n) \in \Xi(N,\delta/d^{n-1})} \#\Sigma(\mathbf{x}_1, \dots, \mathbf{x}_n; N, \delta/d^{n-1}). \quad (2.10)$$

When $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Xi(N, \delta/d^{n-1})$, we get $\mathbf{x}_i \in \Delta(X_i; N, \delta)$ as (2.4) for $1 \le i \le n$. Hence it is seen that

$$\begin{aligned}
\#\Xi(N, \delta/d^{n-1}) &\le \prod_{i=1}^n \#\Delta(X_i; N, \delta) \\
&= \prod_{i=1}^n \#\big\{(N_1, \dots, N_d) : N_l \ge 0 \text{ is an integer in} \\
&\qquad\qquad \big(N(p_{X_i}(t_l) - \delta), N(p_{X_i}(t_l) + \delta)\big) \text{ for } 1 \le l \le d\big\} \\
&< (2N\delta + 1)^{nd}. \qquad (2.11)
\end{aligned}$$

For any fixed $(\mathbf{x}_1, \dots, \mathbf{x}_n) \in \Xi(N, \delta/d^{n-1})$, suppose $([\sigma_1]_{\mathbf{x}_1}, \dots, [\sigma_n]_{\mathbf{x}_n}) \in \Sigma(\mathbf{x}_1, \dots, \mathbf{x}_n; N, \delta/d^{n-1})$; then we get

$$\#\big([\sigma_1]_{\mathbf{x}_1} \times \cdots \times [\sigma_n]_{\mathbf{x}_n}\big) \ge \prod_{i=1}^n \left(\min_{\mathbf{x} \in \Delta(X_i; N, \delta)} \prod_{t \in \mathcal{X}} N_{\mathbf{x}}(t)! \right)$$

similarly to (2.6). Therefore,

$$\begin{aligned}
&\#\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, \delta/d^{n-1}) \\
&\quad \ge \sum_{([\sigma_1]_{\mathbf{x}_1},\dots,[\sigma_n]_{\mathbf{x}_n}) \in \Sigma(\mathbf{x}_1,\dots,\mathbf{x}_n; N, \delta/d^{n-1})} \#\big([\sigma_1]_{\mathbf{x}_1} \times \cdots \times [\sigma_n]_{\mathbf{x}_n}\big) \\
&\quad \ge \#\Sigma(\mathbf{x}_1, \dots, \mathbf{x}_n; N, \delta/d^{n-1}) \cdot \prod_{i=1}^n \left(\min_{\mathbf{x} \in \Delta(X_i; N, \delta)} \prod_{t \in \mathcal{X}} N_{\mathbf{x}}(t)! \right). \qquad (2.12)
\end{aligned}$$

By (2.10)–(2.12) we obtain

$$\#\Delta(X_1, \dots, X_n; N, \delta/d^{n-1}) \le \frac{\#\Delta_{\mathrm{sym}}(X_1, \dots, X_n; N, \delta/d^{n-1}) \cdot (2N\delta + 1)^{nd}}{\prod_{i=1}^n \big(\min_{\mathbf{x} \in \Delta(X_i; N, \delta)} \prod_{t \in \mathcal{X}} N_{\mathbf{x}}(t)! \big)}$$

so that

$$\begin{aligned}\frac{1}{N}&\log \#\Delta(X_1,\dots,X_n;N,\delta/d^{n-1})\\&\le \frac{1}{N}\log \gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym}}(X_1,\dots,X_n;N,\delta/d^{n-1})\bigr)\\&\quad-\sum_{i=1}^n \min_{\mathbf{x}\in\Delta(X_i;N,\delta)}\left(\frac{1}{N}\sum_{t\in\mathcal{X}}\log N_{\mathbf{x}}(t)!\right)+\frac{n}{N}\log N!+\frac{nd}{N}\log(2N\delta+1).\end{aligned}$$

Since it follows similarly to (2.8) that

$$-\frac{1}{N}\sum_{t\in\mathcal{X}}\log N_{\mathbf{x}}(t)!+\frac{1}{N}\log N!\le S(p_{X_i})+\varepsilon+o(1)\quad\text{as } N\to\infty$$

with uniform $o(1)$ for all $\mathbf{x}\in\Delta(X_i;N,\delta)$, we obtain

$$S(p_{(X_1,\dots,X_n)})\le -\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)+\sum_{i=1}^n S(p_{X_i})+n\varepsilon$$

by Lemma 2.1 again, and hence

$$\overline{I}_{\mathrm{sym}}(X_1,\dots,X_n)\le -S(X_1,\dots,X_n)+\sum_{i=1}^n S(X_i). \tag{2.13}$$

The conclusion follows from (2.9) and (2.13). □

In particular, the mutual information $I(X_1\wedge X_2)$ of X_1 and X_2 is equivalently expressed as

$$\begin{aligned}I(X_1\wedge X_2)&=S(p_{(X_1,X_2)},p_{X_1}\otimes p_{X_2})=-S(p_{(X_1,X_2)})+S(p_{X_1})+S(p_{X_2})\\&=I_{\mathrm{sym}}(X_1,X_2)=\overline{I}_{\mathrm{sym}}(X_1,X_2).\end{aligned}$$

Similarly to the problem (2) mentioned in the last of Section 1, it is unknown whether the limit

$$\lim_{N\to\infty}\frac{1}{N}\log \gamma_{S_N}^{\otimes n}\bigl(\Delta_{\mathrm{sym}}(X_1,\dots,X_n;N,\delta)\bigr)$$

exists or not.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Second edition, Wiley-Interscience, Hoboken, NJ, 2006.

[2] I. Csiszár and P. C. Shields, *Information Theory and Statistics: A Tutorial*, in “Foundations and Trends in Communications and Information Theory,” Vol. **1**, No. 4 (2004), 417-528, Now Publishers.

[3] F. Hiai, T. Miyamoto and Y. Ueda, Orbital approach to microstate free entropy, preprint, 2007, math.OA/0702745.

[4] F. Hiai and D. Petz, *The Semicircle Law, Free Random Variables and Entropy*, Mathematical Surveys and Monographs, Vol. 77, Amer. Math. Soc., Providence, 2000.

[5] D. Voiculescu, The analogues of entropy and of Fisher's information measure in free probability theory, II, *Invent. Math.* **118** (1994), 411–440.

[6] D. Voiculescu, The analogue of entropy and of Fisher's information measure in free probability theory VI: Liberation and mutual free information, *Adv. Math.* **146** (1999), 101–166.

Graduate School of Information Sciences, Tohoku University, Aoba-ku, Sendai 980-8579, Japan

Alfréd Rényi Institute of Mathematics, Hungarian Academy of Sciences, H-1053 Budapest, Reáltanoda u. 13-15, Hungary